ADR ENGLAND

Data Explained

Annual Survey of Hours and Earnings linked to Pay-As-You-Earn and Self-Assessment Data – England, Scotland, and Wales

Author: Felix Ritchie, Van Phan, Damian Whittard, Alex Bryson, Anni Caden, John Forth, Rachel Scarfe, Carl Singleton
Date: January 2025

This Data Explained summarises experiences and learning from working with the Annual Survey of Hours and Earnings linked to Pay-As-You-Earn and Self-Assessment Data – England, Scotland, and Wales (ASHE-PAYE-SA) dataset. This publication is intended to help guide future researchers using this data and to provide feedback into future dataset development and documentation.

The administrative data discussed in this Data Explained was made securely available through Wage and Employment Dynamics (WED), funded by ADR UK. The data used in this research project comes from HM Revenue and Customs (HMRC) and the Office for National Statistics (ONS) and was accessed through the ONS Secure Research Service. The data was not originally collected for research and it is expected that there are gaps and inconsistencies in its recording, a number of which are detailed in the following.

Introduction

The data in this collection comes from HMRC's administrative system for personal tax records. The data includes payments made to employees (PAYE), some additional information on employment and end-of-year summaries, and submissions for self-assessment (SA) required from most people who receive non-employment income. These PAYE and SA files can be linked to the ONS Annual Survey of Hours and Earnings (ASHE), a survey aimed at 1% of employees collected by ONS for Great Britain.

The collection comprises seven research-ready datasets compiled from this data:

- weekly_clean: PAYE data for those paid weekly
- monthly_clean: PAYE data for those paid monthly
- weekly_panel: PAYE data for all employees, with monthly pay split across weeks
- monthly_panel: PAYE data for all employees, with weekly pay added up to months
- ASHE_supplement: a summary of the PAYE data for an employee for a year, intended to be used in collaboration with ASHE data
- SA: all the data from the self-assessment records collected into one file
- SA_summary: a subset of the SA data intended to contain the most useful variables

A link field 'hmrc_id' is used throughout the data to provide a consistent reference across the different datasets. The data is for the 1% of the population covered in ASHE (all persons with a National Insurance number ending in a specified two digits) and can be linked directly to ASHE through the 'piden' field.

The PAYE data includes both payments for employment, and occupational pensions. The latter account for about 30% of the total, and are identified with a marker.

How is the data collected?

The data is extracted from HMRC administrative records.

'PAYE' are the submissions made by employers on payments made to employees through HMRC's Real Time Information (RTI) system. Each submission is identified by a pay period (either week 1-52 or month 1-12), with the tax year starting on 6 April. These could involve multiple resubmissions for the same pay period as an employer updates its records, and so the WED files take the last RTI submission in any pay period as the definitive one, so that any employee only has one payment record per PAYE scheme per pay period.

The source data also contains two additional files: one on the details of the employment (start/end date, PAYE scheme), and one containing end-of-year data similar to a P60 form (total pay for the year, maternity pay, student loan repayment and others). These are not currently used, except indirectly to identify occupational pensions and the number of jobs (PAYE scheme number is used as a proxy for employer).

The SA files come from the annual submissions made by individuals with non-employment income, and the variables correspond closely to the fields on the self-assessment form. The data originally comes in multiple files, each one corresponding to a subsection of the tax declaration form; we have combined them all into one record per individual per year.

ASHE is collected in April each year, so 'ASHE year 2016' (collected in April 2016) will correspond to 'tax year 2017' (April 2016-March 2017). ASHE is fully described in the ONS metadata catalogue.

The PAYE data is available from tax year 2015-2019. The SA data is available from tax years 2011-2018. ASHE data is available 1997-2022 (tax years 1998-2023).

The number of distinct individuals in the datasets are detailed in Table 1:

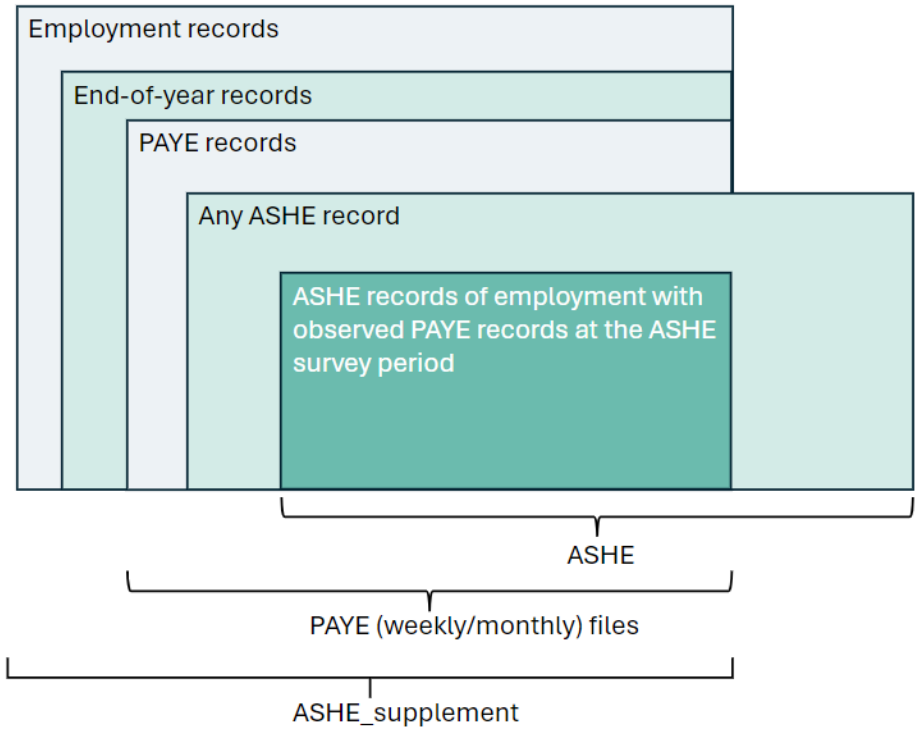
Tax year	weekly paid	monthly paid	Combined (panel)	SA
2011	--	--	--	92,062
2012	--	--	--	91,344
2013	--	--	--	97,119
2014	--	--	--	100,823
2015	102,919	279,327	345,171	102,563
2016	103,856	287,193	352,217	104,243
2017	103,001	293,295	357,826	106,093
2018	102,973	297,975	361,131	104,180
2019	100,154	301,100	361,777	92,062

Table 1 Numbers in the PAYE-SA datasets

The combined totals are smaller than the sum of individual and weekly paid jobs, as some individuals hold both types of job.

Two files, PAYE-employment and PAYE-EoY, provide information on spells of employment and on end-of-year payments. These are available for research but have not been documented yet. They are used in the ASHE_supplement to provide some additional variables. They contain more observations, as an individual may have an employment or end-of-year record, but without any observed pay submissions in a particular year.

In theory, ASHE is a subset of the PAYE data, as the latter should represent the universe of employees. However, there is a small number of individuals with an ASHE record who do not exist anywhere in the HMRC data; this is under investigation, but it suggested to be an administrative or reporting error. The total number of observations is represented in figure 1:



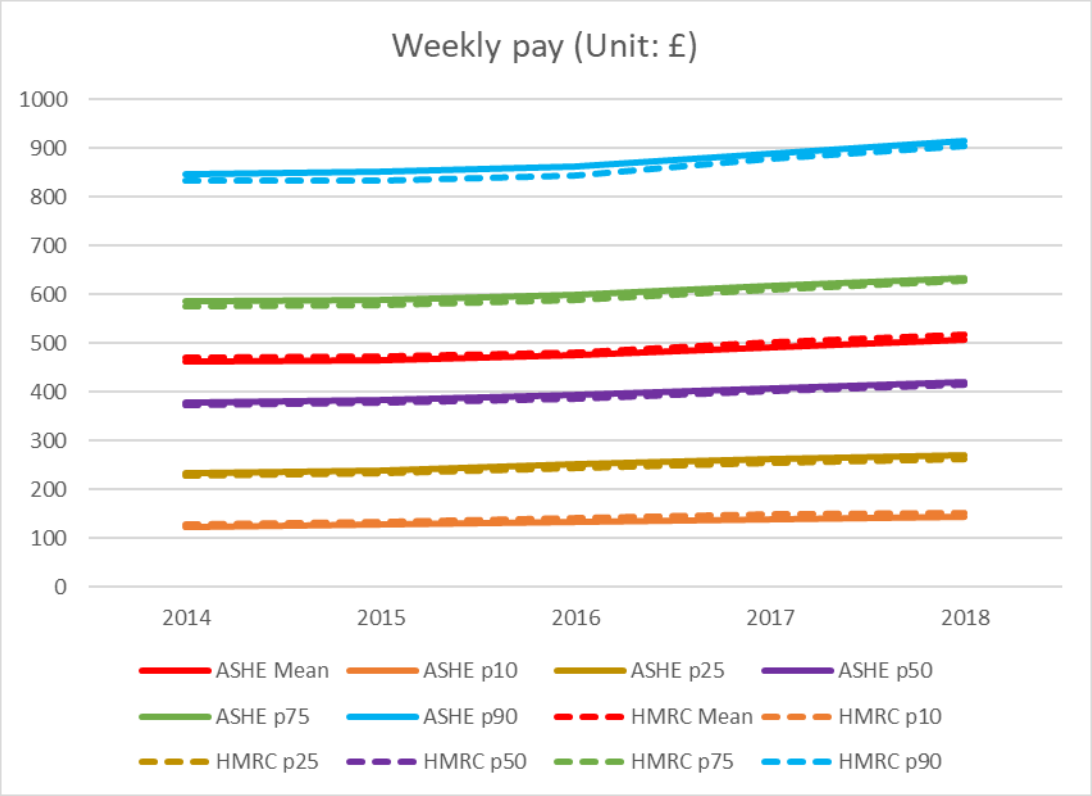
The total number of jobs in the PAYE data, as measured by being part of a separate PAYE index scheme, is higher than the number of individuals. See Table 2:

Tax year	Number of employments in this tax year						
	None	1	2	3	4	5	6+
2015	82,828	243,915	58,301	15,068	4,446	1,421	113,235
2016	87,686	244,374	59,453	15,833	4,709	1,610	103,063
2017	89,232	246,806	59,279	15,907	4,808	1,686	98,578
2018	89,612	245,789	61,197	16,687	5,247	1,846	95,591
2019	91,212	247,241	60,544	16,743	5,414	1,881	96,132

Table 2 Number of employments

ONS estimates that there are roughly 30 million jobs in Great Britain in this period, so these figures show that the dataset holds more than the 1% (300,000 employments) that is in scope for ASHE. These figures include both employment and occupational pensions. This distinction is not directly captured in the RTI system, but HMRC have created an indicator for pensions ('occ_pension'). Following advice from an HMRC-ONS working group on RTI statistics, the WED team have created an enhanced marker ('occ_pension_adj') to take account of additional information. Overall, occupational pensions account for 25% of the PAYE observations, and are overwhelmingly concentrated in those aged 55-plus.

After removing those records which are judged to refer to occupational pension payments, the distribution of wages in the PAYE administrative data is very similar to the distribution in ASHE. The figure below compares PAYE and ASHE data for people who has one job at the point in time of the ASHE survey (April in the tax year). Dotted lines represent HMRC data - ASHE is a solid line.



Mean and median weekly wages are almost identical. ASHE data appears to have slightly more higher-paid employees but the differences are negligible.

The Wage and Employment Dynamics website provides more information in several documents:

- *HMRC quick user guide*
- *PAYE variables list*
- *SA variables list*
- *Creating the PAYE panel*
- *Understanding ASHE and HMRC reference numbers.*

Key variables

There are three key linking variables in the data:

Variable	Present in	Purpose
hmrc_id	All PAYE and SA source files	Unique person identifier across all the HMRC data
Piden	ASHE source files Added to HMRC files	Unique person identifier for ASHE, linkable to hmrc_id via a lookup tale
index_scheme	PAYE files	Identifies PAYE scheme (can be used as a proxy for employer)

Table 3 Linking variables

The HMRC data provided to the WED team does not contain personal characteristics of the individuals. However, ASHE contains age and gender (male/female). 85%-90% of the individuals in the HMRC data have appeared in ASHE at some time, and so for those individuals age and gender has been to the HMRC records.

What can the data be used for?

Files	Use
PAYE pay files *weekly_clean* and *monthly_clean*	These files contain payslip data from the pay period in question, and so are best suited for looking at the distribution, frequency and volatility of monthly/weekly pay.
PAYE pay file *weekly_panel* and *monthly_panel*	These combine weekly and monthly pay and so are best suited for looking at an individual's total earnings and the impact of gaps in employment.
PAYE pay file *ASHE_supplement*	This file contains year-end summaries of the individual's employment and pay history. It is best suited to multivariate analysis of the determinants of pay and employment.
Linked ASHE and *ASHE_supplement*	ASHE holds information on the components of wages, and was designed for analysing low pay. The ASHE supplement allows this to be augmented by wage and employment details between surveys.
SA files	Best suited to understanding how the components of income substitute and complement each other.
Linked SA-PAYE files	Together, these should provide a complete view of an individual's engagement with employment, self-employment and other forms of income, allowing richer work dynamics to be studied.

Existing research or examples of previous research

As this is a new dataset, only working/conference papers have been produced from this dataset so far. The WED website will be updated with publications as they occur.

Data limitations encountered

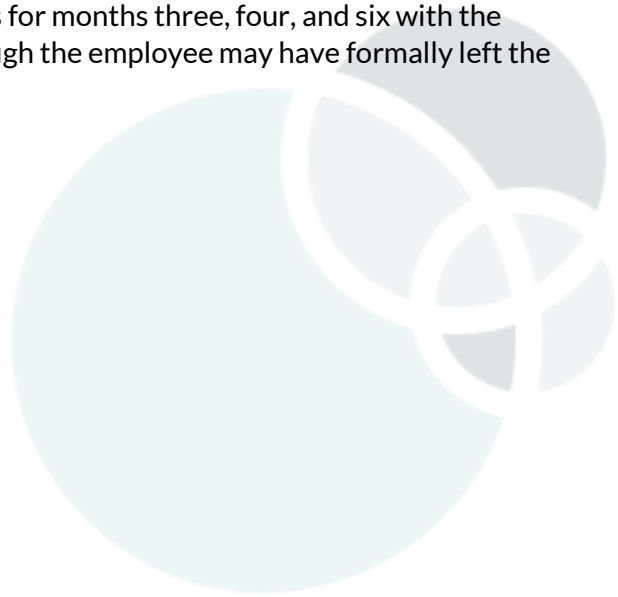
At present, the data is relatively unexplored. The immediate issues with the data (pensions, duplicate RTI submissions, linkage between HMRC files and with ASHE) have been dealt with sufficiently to allow analysis to begin with some confidence.

There are a number of quality checks that could be carried out with the data: for example, consistency between payment dates and employment data, and between regular and end-of-year pay data. This will be carried out over the coming year, but researchers are welcome to investigate and report back, so that this can be added to these data summaries.

One issue that has been noted is that there are a small number of missing pay records which appear to be genuine missing records. For example, if pay is recorded as £100 in week two and three, but cumulative pay shows that this is £100 and £300 in the same weeks. The implication is that week two pay (£100) is missing, not that the person was not paid for the week. The WED team have not adjusted for this as this case is rarely as clear cut as in the example.

There are also cases where pay is recorded as zero. We assume these reflect a continuing employment relationship, where (for example) an employee on a variable-hours contract does no work that week. However, there is no corroborating information to support this, and the assumption is open to criticism.

In the current data delivery there is limited employer information, and nothing that can be linked to the ONS Inter-departmental Business Register, which indexes all ONS business data, including ASHE. As an intermediate step, the PAYE index scheme is taken as a proxy for employer when calculating time with employer or the number of jobs; the time on that scheme is taken as the time with the 'employer'. So, if an employee has pay records for months three, four, and six with the same index scheme, we treat this as one job, even though the employee may have formally left the job in month five and then been re-employed.



Suggested improvements

The most immediate improvements to the dataset will come from these additional variables, in order of importance:

- An enterprise reference linkable to the Business Structure Database, from ONS
- Data on employer or home location from HMRC records
- Age and gender information from HMRC records.

More generally, the time periods of 2014-2019 (PAYE) and 2011-2018 (SA) allowed the WED team to develop an understanding of the dataset, but clearly any long-term use requires regular updates to the dataset.
There are many more variables in the PAYE data, although it is not clear how many of them are derived from the variables already supplied. Ideally all non-derived variables would be made available, except those used for HMRC's internal operations.

Suggested future data linkages

There is substantial value from further onward linking of this data by ONS and HMRC. We suggest that ONS and HMRC prioritise linkage with benefits data from the Department for Work and Pensions, to provide a complete view of an individual's engagement with the labour market family structure, and sources of income.

The second priority for ONS should be to allow linkage with the WED Census-ASHE files, or at least for key variables (ethnicity, disability, country of origin) to be made available. As the index numbers are common across the datasets this should be feasible.

The third priority is to link with the HMRC Migrant Workers Scan. This is already being pursued by the WED team who have been supplied with this data by ONS, with results expected in March 2025.

Recommendations to data owners

The very large number of variables in both SA and PAYE files make it difficult for researchers to identify useful research targets. This is largely because HMRC documentation is derived from administrative processes, rather than a dedicated metadata initiative. We strongly suggest HMRC work with academics (WED, UKDS or others) to develop an improved metadata system which could address both the WED files and the files held in the HMRC Datalab.

Conclusion

The HMC PAYE-SA datasets have enormous research potential, but at present are highly experimental. This Data Explained will be updated as more information about the datasets comes to light from research use.

Disclaimer

This work was produced using administrative data accessed through the ONS Secure Research Service. The use of the data in this work does not imply the endorsement of the ONS or data owners in relation to the interpretation or analysis.

This work uses research datasets which may not exactly reproduce National Statistics aggregates.

Acknowledgements

This work was supported by ADR UK (Administrative Data Research UK). ADR UK is a partnership transforming the way researchers access the UK's wealth of public sector data, to enable better informed policy decisions that improve people's lives. ADR UK is an Economic and Social Research Council (ESRC) investment (part of UK Research and Innovation). [Grant number: ES/W005298/1]

Contact

Name: Felix Ritchie
Email: felix.ritchie@uwe.ac.uk